第 39 卷第 9 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 39 No. 9
Accepted Paper

基于最低度偏置重启随机游走的链路预测方法 *

李巧丽，韩　华†

(武汉理工大学 理学院，武汉 430070)

摘　要：链路预测是数据挖掘主题中的一个重要问题。基于随机游走的相似性方法一般设定游走粒子转移到相邻节点的概率是相等的，忽略了节点度值对转移概率的影响。针对此问题，提出一种基于 lowest-degree 偏置重启随机游走的链路预测方法。首先引入最低度偏置函数，对游走粒子的转移概率进行重新定义，然后将最低度偏置随机游走策略运用到重启随机游走中，探究粒子在游走过程中最低度偏向策略对节点相似度的影响。在九个真实网络数据集上进行链路预测，结果表明，所提方法具有良好的预测精度，且挖掘了更多网络拓扑结构信息，证明该算法在节点相似性的评估上具有一定的优势。

关键词：复杂网络；链路预测；重启随机游走；最低度偏置

中图分类号：TP393;N94　　doi: 10.19734/j.issn.1001-3695.2022.01.0051

Link prediction algorithm based on lowest-degree preference random walk with restart

Li Qiaoli, Han Hua†

(*School of Science*, *Wuhan University of Technology*, *Wuhan* 430070, *China*)

Abstract: Link prediction is an important issue in the subject of data mining. The similarity algorithm based on random walk often set the probability of particles transferring to adjacent nodes to be equal, but ignore the effect of node degree on the transition probability. To solve this problem, this paper proposed a link prediction algorithm based on lowest-degree preference random walk with restart. Firstly, the algorithm redefined the transition probability of the walkers by introduce lowest-degree preference function, then applied it to the random walk with restart, and explored the effect of lowest-degree preference strategy on node similarity. The experimental results of nine real networks show that the proposed method has higher prediction accuracy, and gives more network topology information, which proves that the algorithm has certain advantages in the evaluation of node similarity.

Key words: complex networks; link prediction; random walk with restart; lowest-degree preference

0　引言

近年来，网络科学领域的研究蓬勃发展，越来越多的复杂系统成为复杂网络学[1]的研究对象。复杂系统中的个体和个体间的联系可以抽象为复杂网络来表示。常见的复杂网络有生物网络[2]、社会网络[3]、通信网络[4]等。链路预测作为复杂网络的重要研究工具，旨在借助网络中已知数据信息挖掘网络中未知的连边关系[5~7]。链路预测的研究在众多领域发挥着重要价值，从理论上来说，可以帮助更好的理解网络演化机制及网络动力学行为[8]；从应用上来说，当前社交网络上的用户拓展、电信网络上的诈骗源头识别、电商网络上的客户精准营销等[9, 10]都是链路预测在现实网络中的典型应用。

目前，许多经典的链路预测算法被提出。基于相似性的链路预测算法应用领域最为广泛。基于网络结构相似性的方法可大致上分为：a)基于局部信息的方法[5]；b)基于路径的相似性方法[11]；c)基于随机游走的方法[12]。基于局部信息的方法主要利用节点的局部信息(如节点的度、共同邻居数目等)进行链路预测。这类方法的计算复杂度较低，但往往以牺牲精度为代价。基于路径的方法倾向于利用节点之间的路径信息(如节点之间路径数量，路径中间节点的信息等)计算节点相似性。这类方法在涉及到多阶路径信息以及全局路径信息时，计算复杂度相对较高。基于随机游走的方法是基于粒子随机游走过程定义的，即假设粒子从初始节点开始，以一定的概率随机游走到它的相邻节点，这个过程一直持续到粒子出现在每个节点上的概率分布达到平稳状态。这类指标只关注节点邻居的局部信息，可以在计算复杂度和预测性能之间取得良好的折中，因此还被广泛应用于推荐系统、信息传播和社团划分等问题中[13, 14]。

随机游走的这一优势使其成为解决链路预测问题的主要方法，并因此取得了许多成果。一个典型的例子是 PageRank[15]算法，其中随机游走方法起着关键作用。此外，Li 等人[16]认为在现实网络中，节点不仅趋向于连接度小的节点，而且也趋向于连接中心节点，提出一种最大熵随机游走的链路预测算法，此算法涉及到网络节点中心性的计算，复杂度相对较高。文献[17]通过 deepwalk 网络表示学习算法得到节点的向量表示，并通过欧氏距离表征各节点的结构相似度，提出一种网络表示学习与随机游走的链路预测算法，该算法在预测过程中同时考虑网络结构信息和节点属性信息，在处理较大规模的网络时很吃力。Jin 等人[18]提出了一种有监督和扩展的重启随机游走方法，其中每个节点对应一个重启概率，实验结果表明，所提算法为排名和链接预测任务提供了较好性能，但节点重启概率的设置具有非普适性，限制了该类算法的应用范围。

上述基于随机游走的方法大多数使用均匀分布来定义粒子的转移概率，忽略了节点局部区域的细微结构对转移概率的影响[19~21]。事实上，由网络的度度相关性[22]可以看出，节

收稿日期：2022-01-20；**修回日期：**2022-03-15　**基金项目：**国家自然科学基金青年科学基金资助项目(111701435)；国家自然科学基金资助项目(12071364)

作者简介：李巧丽(1991-)，女，河南平舆人，硕士研究生，主要研究方向为复杂网络动力学、链路预测；韩华(1975)，女(通信作者)，山东莱州人，教授，博士，主要研究方向为复杂性分析与评价、经济控制与决策(1104768792@qq.com).

点之间的连接不是随机产生的，粒子在游走过程中会受到节点度值的影响。最近，文献[23]发现，随机游走者通常频繁访问网络上的高度节点，这种搜索策略更有可能导致较低的搜索效率，并受 PageRank 算法[24]的启发，提出一种最低度偏好随机游走的搜索策略(LPRW)，实验结果表明，与无偏向的随机游走相比，LPRW 方法可以显著减少搜索时间。吕等[25]人认为粒子在游走过程中具有一定的度偏向性，提出了 BRWR 方法，实验结果同样表明，粒子偏向游走到高度节点的程度越大，预测的精度越低。

受上述方法和 PageRank 算法的启发，本文提出了一种最低度偏置重启随机游走链路预测算法，该算法是由纯随机游走策略和仅访问最低度邻居组成的混合游走策略，并将其应用到链路预测中。该方法首先通过引入最低度偏置函数，对游走粒子的转移概率进行重新定义；然后将最低度随机游走策略运用到重启随机游走中，探究粒子在游走过程中最低度偏向策略对其转移的作用；最后通过多个真实网络数据集验证了所提方法的有效性。

1 相关工作

1.1 问题描述

给定一个无权无向网络，用一个二元序对 $G=(V,E)$ 表示，包含 $|V|=N$ 个节点和 $|E|=M$ 条边。对于网络中所有的节点，所有可能产生连边的两点集合用 $\Omega=V\times V$ 表示。连通的网络 G 可以用邻接矩阵 $\boldsymbol{A}=(a_{uv})_{N\times N}(u,v\in V)$ 表示，其中 $\boldsymbol{A}$ 中的元素 $a_{uv}=1$，则代表节点对 (u,v) 之间有连边，否则 $a_{uv}=0$。预测算法为网络中每一对未连接的节点赋予一个相似性分数值 S_{xy}。将所有 S_{xy} 降序排列，排在最前面的边存在的可能性越大。

在实际预测中，一般根据不同评价需求设定相似分数阈值，相似度高于阈值的连边将选取为推荐结果；或根据相似分数值排序结果，选取前面 l 条预测连边作为预测结果。预测连边进一步可应用于电商推荐系统或在生物实验中作为指导依据等。

1.2 链路预测方法

对于网络中任意两个节点 $u,v\in V$，设 $\Gamma(u)$ 和 $\Gamma(v)$ 分别为节点的邻居集合，以 $|\Gamma(u)|$ 表示集合的势，$\Gamma(u)\cap\Gamma(v)$ 表示节点的共同邻居集合，k_u 代表节点的度。下文对几种常用的相似性指标[7]介绍如下：

a)共同邻居(CN)。通过节点对之间的共邻节点的个数刻画节点 u 和 v 的相似性，用(1)式表示为

$$S_{uv}^{\mathrm{CN}}=|\Gamma(u)\cap\Gamma(v)| \tag{1}$$

其中，$\Gamma(u)$ 为节点 u 的邻居集合，$|\ |$ 表示集合的势。

b)PA 指标。基于节点间的偏好连接特性提出的指标，认为节点更倾向于与高度节点相连，即

$$S_{uv}^{\mathrm{PA}}=k_uk_v \tag{2}$$

c)RA 指标。是一种基于共享特征的相似性度量方法，其思想是度小的共邻节点的贡献大于度大的共邻节点，采用共邻节点的度的倒数对相似性进行加权，则节点的相似性定义为

$$S_{uv}^{\mathrm{RA}}=\sum_{\omega\in\Gamma(u)\cap\Gamma(v)}\frac{1}{k_\omega} \tag{3}$$

d)HDI 指标。该指标称为高度节点不利指标，

$$S_{uv}^{\mathrm{HDI}}=\frac{|\Gamma(u)\cap\Gamma(v)|}{\max\{k_u,k_v\}} \tag{4}$$

e)Katz 指标。该指标实际上是一种最短路径方法，考虑了两个节点间所有跳的路径数，并根据路径长度的不同采取分级惩罚，即

$$S_{uv}^{\mathrm{Katz}}=\sum_{l=1}^{\infty}\beta^l\,|path_{u,v}^{<l>}|=\beta\boldsymbol{A}_{uv}+\beta^2(\boldsymbol{A}^2)_{uv}+\beta^3(\boldsymbol{A}^3)_{uv}+\cdots \tag{5}$$

其中，β 为路径权重调节参数，$|path_{u,v}^{<l>}|$ 代表连接节点 u 和 v 之间路径长度为 l 的路径数。

f)SimRank 指标(SimR)。它假设如果两个节点所相连的节点相似，则这两个节点就相似，描述了两个分别从节点 u 和节点 v 出发的粒子相遇时平均经过的时间。用式(6)表示为

$$S_{uv}^{\mathrm{SimR}}=C\frac{\sum_{\omega\in\Gamma(u)}\sum_{\omega'\in\Gamma(v)}S_{\omega\omega'}^{\mathrm{SimR}}}{k_uk_v} \tag{6}$$

其中，假定 $S_{uu}=1$，$C\in[0,1]$ 代表相似性传递时的衰减参数。

g)平均通勤时间(ACT)。基于随机游走定义的相似性指标，表示一个粒子从节点 u 游走到节点 v 所需走的平均步数，则节点的相似性表示为

$$S_{uv}^{\mathrm{ACT}}=\frac{1}{l_{uu}^{+}+l_{vv}^{+}-2l_{uv}^{+}} \tag{7}$$

其中，l_{uv}^{+} 代表网络的拉普拉斯矩阵中第 u 行第 v 列对应的元素值。

h)有重启的随机游走指标(RWR)。该指标是由 PageRank 算法拓展而来的。它是指执行随机游走的粒子在每走一步都可能以一定概率返回到它的初始位置。设粒子返回概率为 $1-c$，网络的马尔可夫转移矩阵 $\boldsymbol{P}$ 可表示为 $p_{uv}=a_{uv}/k_u$，其中 p_{uv} 和 a_{uv} 分别表示矩阵 $\boldsymbol{P}$ 和邻接矩阵 $\boldsymbol{A}$ 中的元素。某一个粒子初始时刻在节点 u，则 $t+1$ 时刻到达网络中各个节点的概率分布向量可表示为

$$\boldsymbol{\pi}_u(t+1)=c\cdot\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\pi}_u(t)+(1-c)\boldsymbol{e}_u \tag{8}$$

其中，$\boldsymbol{e}_u$ 代表初始状态。上式的稳定解可以表示为 $\boldsymbol{\pi}_u=(1-c)(\boldsymbol{I}-c\boldsymbol{P}^{\mathrm{T}})^{-1}\boldsymbol{e}_u$，其中 $\boldsymbol{\pi}_u$ 代表稳态解向量，π_{uv} 代表 π_u 的第 v 个元素，则 RWR 相似性定义为

$$S_{uv}^{\mathrm{RWR}}=\pi_{uv}+\pi_{vu} \tag{9}$$

2 基于最低度偏置重启随机游走的相似性方法

随机游走在复杂网络领域中起着至关重要的作用，并在各个领域取得了一系列研究成果，包括社区检测、链接预测、重要节点挖掘等，一般分为纯随机游走和有偏随机游走[26]。纯随机游走是指游走者从任意节点或源节点 u 开始，只能以等概率随机游走的方式跳到一个相邻节点。相比之下，有偏随机游走是指在未知网络中强制寻找最近的目标节点进行游走。一个有偏向的随机游走者从当前节点跳转到潜在的新节点之一的跳转概率是不等的，并且游走者倾向于访问倾向于访问或忽略高拓扑属性值的节点，包括强度，集聚系数或度等。因此，本文假设粒子在随机游走的过程中，采用纯随机游走和偏向于访问最低度邻居的混合游走策略，并基于混合游走策略得到粒子的跳转概率矩阵。在此基础上，让粒子以重启随机游走的方式进行游走，对网络中未连边的节点对进行相似性计算，找到每个网络最佳的最低度偏置调节参数，以达到提高预测精度的目的。

2.1 最低度偏置的重启随机游走

定义 1 最低度偏置转移概率。考虑一个在网络相邻节点之间跳跃的粒子，由马尔可夫过程[27]可知，粒子下一个时刻的状态只与现在的状态有关。基于最低度偏置随机游走过程中，在每一个时间步，游走者采取纯随机游走和偏向于访问最低度邻居节点的混合游走策略，使用一个可变参数 β 来调整两者的融合比率，则当前在节点 u 的游走者跳转到节点 v 的转移概率[23]定义如下：

$$w_{uv}=(1-\beta)w_{uv}^{(1)}+\beta w_{uv}^{(2)} \tag{10}$$

其中，$\beta\in(0,1)$，$w_{uv}^{(1)}=a_{uv}/k_u$ 表示纯随机游走策略的转移概率，$w_{uv}^{(2)}$ 表示最低度游走策略的概率。$w_{uv}^{(2)}$ 的定义如下

$$w_{uv}^{(2)}=\begin{cases}\dfrac{1}{\mathrm{card}(U_v)}, & v\in U_v\\ 0, & v\notin U_v\end{cases} \tag{11}$$

其中，U_v 表示节点 u 的最低度邻居节点的集合，$\mathrm{card}(U_v)$ 表示最低度邻居节点的个数。值得注意的是，当 $\beta=0$ 时，最低度偏好随机游走退化为通用随机游走，这种情况下，游走者在任何时间停留在节点 u 上的平稳状态概率与节点 u 的度数成正比[27]，因此游走者更有可能在搜索过程中访问度数高的节点。而在最低度偏好随机游走的过程中游走者同时采取 $\beta>0$ 时的最低度搜索策略，因此避免了这种情况的发生。图 1 给出了 $\beta=1/3$ 时最低度偏置随机游走的转移概率。

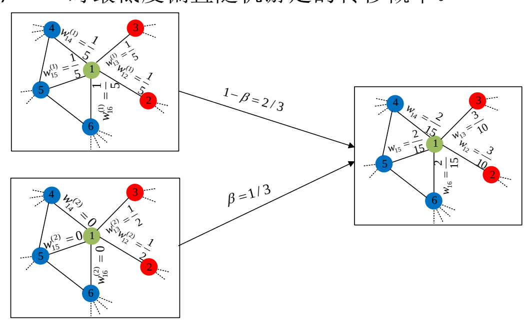


图 1　最低度偏置随机游走示意图

Fig. 1　Schematic diagram of the least lowest-degree preference random walk

基于最低度偏置的重启随机游走是指游走粒子从网络中的某一个节点出发，每游走一步将选择是以概率 $1-\alpha$ 跳转到相邻节点，还是以概率 α 返回初始位置。如果粒子选择以概率 $1-\alpha$ 跳转到相邻节点，此时会以定义 1 中的最低度偏置转移概率 w_{uv} 选择下一步跳转到的节点，重复上述过程，直到达到平稳状态。采用最低度偏置转移的重启随机游走既避免了随机游走过程中节点等概率转移和偏向高度节点游走的现象，又解决了有偏置随机游走在达到平稳状态之前游走粒子就发生终止的问题。

定义 2　最低度偏置重启随机游走指标。将定义 1 中节点的最低度偏置转移概率用于有重启随机游走中进而得到最低度偏置的重启随机游走算法(lowest-degree preference random walk with restart，简称 LPRWR 算法)。令 $\pi_{uv}(t)$ 表示粒子在时间 $t=0$ 从节点 u 出发，在 t 时刻停留在节点 v 的概率。这个概率的演化由下面的主方程给出，定义为

$$\pi_{uv}(t+1)=(1-\alpha)\sum_{l=1}^{N}a_{lv}w_{lv}\pi_{ul}(t)+\alpha\pi_{uv}(0) \tag{12}$$

其中，α 为重启概率，$\pi_{uv}(0)$ 表示初始状态向量的的第 v 个元素。

令一步转移概率的矩阵表示为 $\boldsymbol{W}$，则随机游走的迭代公式定义为

$$\boldsymbol{\pi}_u(t+1)=(1-\alpha)\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\pi}_u(t)+\alpha\boldsymbol{\pi}_u(0) \tag{13}$$

根据 C-K 方程，粒子的 m 步转移概率可表示为 $(\boldsymbol{W}^{T})^{m}$，所以粒子随机游走 m 步的迭代公式为

$$\boldsymbol{\pi}_u(t+m)=(1-\alpha)(\boldsymbol{W}^{\mathrm{T}})^{m}\boldsymbol{\pi}_u(t)+\alpha\boldsymbol{\pi}_u(0) \tag{14}$$

当 $t\to\infty$ 时，由马尔可夫链的平稳状态[26]可知，随机游走的概率分布可能会收敛到一个极限概率分布，也既是平稳分布，即满足 $\boldsymbol{\Pi}=(1-\alpha)\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Pi}+\alpha\boldsymbol{\pi}_u(0)$，因此式(14)可以改写为

$$\begin{aligned}\boldsymbol{\pi}_u&=(1-\alpha)\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\pi}_u+\alpha\boldsymbol{\pi}_u(0)\\&=\alpha(\boldsymbol{I}-(1-\alpha)\boldsymbol{W}^{\mathrm{T}})^{-1}\boldsymbol{\pi}_u(0)\\&=R\boldsymbol{\pi}_u(0)\end{aligned} \tag{15}$$

其中，$\boldsymbol{\pi}_u$ 为稳态时的概率分布；R 为初始状态 $\boldsymbol{\pi}_u(0)$ 下的点的相关度。计算稳态解时所有路径都已考虑。R 可写成无穷级数的形式：

$$R=\alpha(\boldsymbol{I}-(1-\alpha)\boldsymbol{W}^{\mathrm{T}})^{-1}=\alpha\sum_{n=0}^{\infty}(1-\alpha)^{n}(\boldsymbol{W}^{\mathrm{T}})^{n} \tag{16}$$

上式中，R 还可以看做 $(\boldsymbol{W}^{\mathrm{T}})^{n}$ 的加权和，其元素 $(\boldsymbol{W}^{\mathrm{T}})^{n}_{uv}$ 表示经过 n 次迭代后，随机游走粒子从节点 u 停留在节点 v 的概率。n 表示一个大规模的转换，随着 n 的不断增加，随机游走将转换的更远[28]。故 LPRWR 算法可以认为是基于考虑两节点之间转移的所有路径来对相似性进行优化。由此定义 LPRWR 相似度为

$$S_{uv}^{\mathrm{LPRWR}}=\pi_{uv}+\pi_{vu} \tag{17}$$

其中，元素 π_{uv} 代表由节点 u 出发的粒子最终到达节点 v 的概率。

综上所述，该算法的流程如下：

算法 1　LPRWR 算法

输入：网络邻接矩阵 $\boldsymbol{A}=(a_{uv})_{N\times N}(u,v\in V)$，最低度偏置调节参数 β，重启概率 α。

输出：网络的节点相似度得分矩阵 $\boldsymbol{S}$。

a) 初始化最低度偏向转移矩阵 $\boldsymbol{W}\leftarrow\boldsymbol{O}_{N\times N}$，节点相似度得分矩阵 $\boldsymbol{S}\leftarrow\boldsymbol{I}_{N\times N}$

b) for i=1 to N，j=1 to N

c)　根据式 $w_{uv}=(1-\beta)w_{uv}^{(1)}+\beta w_{uv}^{(2)}$ 计算节点间的最低度偏置转移概率；更新最低度偏置转移矩阵 $\boldsymbol{W}$

d)　for i=1 to N do

e)　　$\pi_u=\alpha(\boldsymbol{I}-(1-\alpha)\boldsymbol{W}^{\mathrm{T}})^{-1}\boldsymbol{\pi}_u(0)$　/* 计算节点 u 和网络中其他各节点的相似度得分值 */

f)　End While

g) End for

h) Return $\boldsymbol{S}$

2.2　算法收敛性

LPRWR 算法中粒子随机游走过程的收敛性是保证算法能应用的必要条件，下文给出算法收敛性的严格证明。

定理 1　LPRWR 算法是收敛的。

证明： a)由于最低度偏置转移矩阵 $\boldsymbol{W}$ 中的元素 w_{uv} 满足 $w_{uv}\ge 0$，$\sum_{v\in V}w_{uv}=1$，$u,v\in V$，因此矩阵 $\boldsymbol{W}$ 是随机矩阵。由随机矩阵性质可得出，矩阵 $\boldsymbol{W}$ 是不可约的。b)随机游走过程是一个马尔可夫链，对于其中的任一状态，当随机游走经过这一状态后，由于存在重启概率，再次遍历这一状态所需游走的步数是不确定的，因此整个游走过程是非周期性的。

由此可得出 LPRWR 算法采用的随机游走过程是各态历经的[29]，故 LPRWR 算法是收敛的。

2.3　复杂度分析

定理 2　LPRWR 算法的时间复杂度是 $O(N^3)$。

证明： 由于在 $t\to\infty$，LPRWR 算法的概率分布会收敛到一个平稳分布，根据稳态解 $\pi_u=\alpha(\boldsymbol{I}-(1-\alpha)\boldsymbol{W}^{\mathrm{T}})^{-1}\boldsymbol{\pi}_u(0)$，故算法的关键是计算矩阵 $(\boldsymbol{I}-(1-\alpha)\boldsymbol{W}^{\mathrm{T}})^{-1}$ 的逆，而求一个 $N\times N$ 矩阵的逆或伪逆的复杂度是 $O(N^3)$，故 LPRWR 算法的时间复杂度是 $O(N^3)$。

3　实验条件介绍

实验中，将网络连边 E 划分为训练集 E^T 和测试集 E^P，其中 $E=E^T\cup E^P$，且 $E^T\cap E^P=\varnothing$。训练集被认为已知信息用于计算未连边节点对的得分，有效的算法应当赋予测试集更高的分值，而对不存在的连边赋予较低的分值。

文中采用十折交叉检验来测试所提算法的性能，并且为了方便进行数据处理，将所有数据以 CSV 格式保存在 MySQL 数据库中。使用 Rapidminer 数据挖掘工具按比例 $E^P:E^T=1:9$ 随机选取训练集和测试集。实验中，每个 AUC 和 Precision 均为不少于 100 次独立实验结果的均值。

3.1　衡量指标

链路预测算法的主流衡量指标包括 AUC(area under the curve)[30]和精确度(Precision)[31]。前者侧重于从整理上评价算法对未知对象的区分度；后者侧重于精准预测，关注的是预测前列结果命中的比率。

AUC 是指在衡量算法性能时，从测试集 E^P 中随机选择一条边的分数值大于一条不存在边的分数值的概率。实验时，若测试集中边的预测分数值大于不存在边的分数值加 1，此种情况次数记为 n' 次，二者相等时则加 0.5，情况次数记为 n'' 次，则 AUC 指标可以表示为

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{18}$$

其中，n 为独立比较的次数，显然，随机预测下 AUC≈0.5。此外，在计算 AUC 时还需考虑到比较次数 n 的取值问题。吕琳媛等[7]证明了：无论测试集比例取何值，n 最多取 672400 次时，能够以 90%的置信度确保 AUC 的绝对计算误差不超过 1‰。因此，在本文实验中 n 均取 672400 次。

Precision 指标关注的是排在前 L 个预测边中预测准确的比率，表示为

$$\text{Precision} = \frac{l}{L} \tag{19}$$

其中，l 代表预测分数值排在前 L 个的连边中出现在 E^P 中的个数。

3.2 数据集

实验选取 9 个不同规模的真实网络数据集，这些数据集均来源于网络公开数据库[32]。包括 Dolphins，Neural，Polbook，Metabolic，Netscience(NS)，Football，Circuit，Facebook，Hamster。上述网络数据集的相关统计特性如表 1 所列。其中，N 与 M 分别为节点数与边数，$\langle k\rangle$ 为网络平均度，$\langle d\rangle$ 为平均最短路径，r 为同配性系数，H 为度异质性，C 为集聚系数。

表 1　9 个真实网络的拓扑特征

Tab. 1　Topological parameters of nine real networks

Network	N	M	$\langle k\rangle$	$\langle d\rangle$	r	H	C
Dolphins	62	159	5.129	3.357	-0.044	1.327	0.259
Neural	297	2148	14.465	2.455	-0.163	1.801	0.308
Polbook	105	441	8.400	3.079	-0.128	1.421	0.488
Metabolic	453	2025	8.940	2.676	-0.226	4.485	0.647
NS	1589	2742	3.451	5.823	0.461	2.010	0.878
Football	115	613	10.66	2.510	0.162	1.690	0.407
Circuit	512	819	3.199	6.858	-0.030	1.259	0.055
Facebook	2888	2981	2.064	3.870	-0.668	0.250	0.003
Hamster	1858	12534	13.490	3.390	-0.085	3.360	0.090

4　实验结果与分析

为了评估 LPRW 方法的性能，本文将首先计算节点间的相似度得分，然后使用 AUC 和 Precision 两个衡量指标来量化本文方法进行链接预测的准确性。在实验中，按照基于随机游走方法中的典型做法，设置重启系数 $\alpha=0.15$[12, 15, 23]。由于篇幅所限，下文只给出 AUC 指标的运行结果。

4.1　相关参数对 AUC 结果的影响

在式(10)中 β 主要用来调节最低度偏置游走的比例，其中 $\beta\in[0,1)$。本文研究了参数 β 对预测结果的影响，实验结果如图 2 所示。结果表明，相比 $\beta=0$(无偏向随机游走)，指标的预测精度都得到一定的提高，且在一定的参数范围内均可以取得最佳的预测精度，这说明最低度偏置游走对相似性的影响是不可或缺的。从图 2 中的每个子图可以观察到，不同网络的 AUC 曲线到达峰值后会呈现不同程度的下降，其中大部分网络如 Dolphins，Neural，Polbook 等网络的下降趋势较快。这在一定程度上表明最低度偏置程度较小时，预测的准确度较高。从图中可以看出，在 Dolphins 网络，Metabolic 网络，NS 网络中，β 在 0.05 时预测效果最好；在 Neural 网络，Hamster 网络中，β 在 0.1 时预测效果更好；对于 Polbook 网络，Facebook 网络，最优的 β 为 0.15；对于 Football 网络，最优的 β 为 0.25；Circuit 网络中，最优的 β 主要分布在 $\beta=0.45$ 和 $\beta=0.1$ 附近。因此，不同的网络取得最优 AUC 值时对应的参数值有一定的不同，然而最优的参数值取得较小时比如在 0 到 0.2 之间时，可以取得较好的预测效果。此外，AUC 值取得最优时 $\beta\neq0$ 也相当于粒子在游走时偏向于度小的节点，这与 RA 指标的思想一致，即低度值的共邻节点的作用大于高度值的共邻节点的作用。综上分析，在实际应用中，可以选取较小的 β 值进行预测。

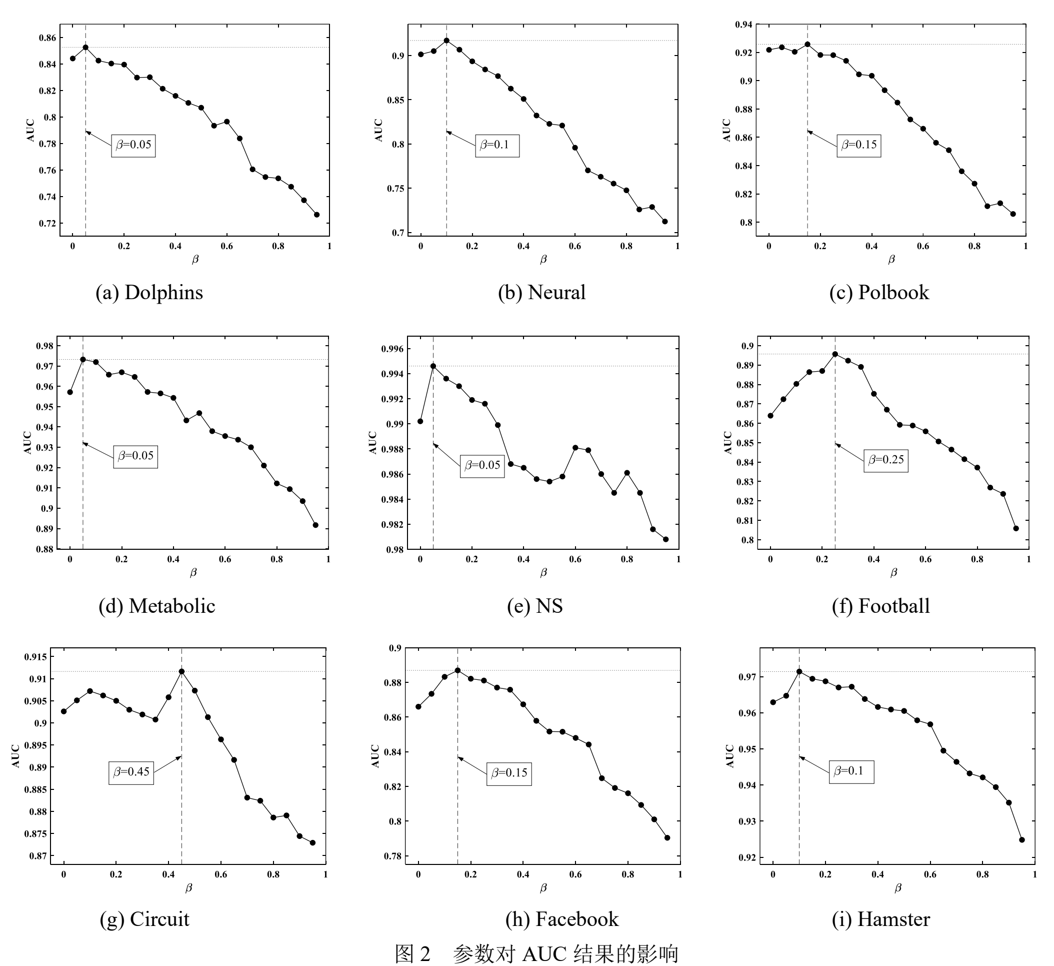


图 2　参数对 AUC 结果的影响

Fig. 2　The influence of the parameter value on the AUC results

4.2 可行性分析

为了进一步验证最低度偏置随机游走的可行性及LPRWR算法的有效性，将所提方法与8个主流指标(包括4个局部指标和4个全局指标)进行预测性能的对比分析，各个指标的AUC值如表2所示。可以看出，LPRWR算法在8个网络中取得了最高 AUC 值，仅在 Facebook 网络中略低于RWR指标。另外，虽然其他几种方法在某些网络上的得分可能接近本文方法，但它们在其他一些网络上的表现存在明显差异。这一事实表明了所提方法预测结果较为稳定，在广泛的网络上具有一定的优势，而其他基准指标可能仅在某些特定的网络上表现良好。

此外，CN、PA、RA、HDI这些局部指标中，RA指标对度大的节点进行惩罚，在局部指标中预测效果相对较好。在Katz、SimR、ACT、RWR这些全局指标中，其中Katz指标是考虑节点之间的所有路径，SimR、ACT、RWR都是基于随机游走过程定义的指标，且 RWR 在这几个指标中整体表现相对较好。若以RWR指标为基准，LPRWR算法预测准确度平均提升了 2.14%，且在 Football 网络中 AUC 结果提升了4.48%。由定理 2 可知 LPRWR 算法的时间复杂度和R W R算法相同，均为$O(N^3)$，在两者时间复杂度相同的清况下，LPRWR方法的预测准确度比RWR方法更好，进一步表明了最低度偏置重启随机游走对链路预测是有效和可行的。

表 2　不同指标下 AUC 结果对比

Tab. 2　Comparison of AUC for different indices

Network	CN	PA	RA	HDI	Katz	SimR	ACT	RWR	LPRWR
Dolphins	0.7666	0.6606	0.7668	0.7948	0.8314	0.8234	0.7713	0.8264	**0.8525**
Neural	0.8465	0.7542	0.8685	0.7784	0.8556	0.7632	0.7425	0.8919	**0.9171**
Polbook	0.8882	0.6707	0.8997	0.8596	0.9005	0.8676	0.7488	0.9080	**0.9258**
Metabolic	0.9240	0.8250	0.9608	0.7638	0.9221	0.7639	0.7700	0.9492	**0.9732**
NS	0.9754	0.6579	0.9814	0.9753	0.9856	0.9819	0.9334	0.9771	**0.9946**
Football	0.8395	0.2950	0.8403	0.8505	0.8511	0.8772	0.5808	0.8573	**0.8957**
Circuit	0.5446	0.4212	0.5448	0.5612	08301	0.9001	0.6747	0.9066	**0.9116**
Facebook	0.8434	0.7514	0.8592	0.7874	0.8421	0.7698	0.7754	**0.8918**	0.8869
Hamster	0.7949	0.8871	0.8065	0.7993	0.9349	0.8390	0.8695	0.9476	**0.9714**

5 结束语

准确预测复杂网络中节点间的相似性对于加快积极信息在网络中传播、预防电信诈骗、促进电商网络的发展具有现实意义。对于当前基于随机游走过程的链接预测方法，大都认为粒子转移到其不同邻居的概率相等，然而，该方法在分析中忽略了网络的详细结构信息。在本文中，通过考虑最低度偏置游走对粒子转移概率的影响，定义了最低度偏置函数，提出一种混合游走策略，并将其应用到重启随机游走中，进而量化节点间的相似性。以提出的方法为基础，在真实网络上经过大量实验，并对各指标的预测效果进行对比分析，证实了所提方法的有效性和可行性，表明该算法在节点相似性的度量上中具有一定的优势。

本文所提算法仅适用于无权无向的单层网络，具有一定的局限性，如何设计适用于加权有向的多层网络的链路预测算法，是接下来要研究的问题。在下一步的研究中，可以尝试挖掘更多的影响随机游走过程的结构信息，将此应用在多层网络上，进一步提高链路预测的准确度。

参考文献：

[1] Tan Yangxin, Wu Junlin, Zhong Qing. Complex network [J]. Journal of Physics: Conference Series, 2020, 1601: 032011.

[2] Cannistra C V, Alanislobato G, Ravasi T. From link prediction in Brain connectomes and protein interactomes to the local community paradigm in complex networks [J]. Scientific Reports, 2013, 3 (4): 1-13.

[3] Fan Tongrang, Xiong Shixun, Zhao Wenbin, *et al.* Information spread link prediction through multi-layer of social network based on trusted central nodes [J]. Peer-to-Peer Networking and Applications, 2019, 12 (5): 1028–1040.

[4] Dzaferagic M, Kaminski N, Mcbride N, *et al.* A functional complexity framework for the analysis of telecommunication networks [J]. Journal of Complex Networks, 2018, 6 (6): 971–988.

[5] Zhou Tao, Lyu Linyuan, Zhang Yicheng. Predicting missing links via local information [J]. European Physical Journal B, 2009, 71 (4): 623-630.

[6] Gul H, Amin A, Adnan A, *et al.* A systematic analysis of link prediction in complex network [J]. IEEE Access, 2021, 9: 20531-20541.

[7] Lyu Linyuan, Zhou Tao. Link prediction in complex networks: a survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390 (6): 1150-1170.

[8] 谭索怡，祁明泽，吴俊，等. 复杂网络链路可预测性：基于特征谱视角 [J]. 物理学报, 2020, 69 (8): 188-197. (Tan Suoyi, Qi Mingze, Wu Jun, *et al.* Link predictability of complex network from spectrum perspective [J]. Acta Physica Sinica, 2020, 69 (8): 188-197.)

[9] Assouli N, Benahmed K, Gasbaoui B. How to predict crime informatics-inspired approach from link prediction [J]. Physica A: Statistical Mechanics and its Applications, 2021 (8): 125-143.

[10] Ai Jun, Liu Yayun, Su Zhan, *et al.* Link prediction in recommender systems based on multi-factor network modeling and community detection [J]. Europhysics Letters, 2019, 126 (3): 38003.

[11] Lyu Linyuan, Jin Cihuang, Zhou Tao. Similarity index based on local paths for link prediction of complex networks [J]. Physical Review E, 2009, 80 (4): 046122.

[12] Tong Hanghang, Faloutsos C, Pan Jiayu, *et al.* Fast random walk with restart and its applications [C]// Proc of the Sixth International Conference on Data Mining. Piscataway: IEEE Press, 2006: 613-622.

[13] Fu Xianghua, Wang Chao, Wang Zhiqiang. Scalable community discovery based on threshold random walk [J]. Journal of Computational Information Systems, 2012, 8 (21): 8953–8960.

[14] 赵海燕，张健，曹健. 基于主题分组与随机游走的App推荐算法 [J]. 计算机应用研究, 2018, 35 (08): 2277-2280. (Zhao Haiyan, Zhang Jian, Cao Jian. Personalized App recommendation algorithm based on topic grouping and random walk [J]. Application Research of Computers, 2018, 35 (08): 2277-2280.)

[15] Nassar H, Benson A R, Gleich D F. Neighborhood and PageRank methods for pairwise link prediction [J]. Social Network Analysis and

Mining, 2020, 10 (1): 63.

[16] Li Ronghua, Yu Jeffreyxu, Liu Jianquan. Link prediction: the power of maximal entropy random walk [C]// Proc of the 20th ACM Conference on Information and Knowledge Management. United Kingdom: ACM Press, 2011: 24-28.

[17] 刘思，刘海，陈启买，等. 基于网络表示学习与随机游走的链路预测算法 [J]. 计算机应用, 2017, 37 (8): 2234-2239. (Liu si, Liu Hai, Chen Qimai, *et al.* Link prediction algorithm based on network representation learing and random walk [J]. Journal of Computer Applications, 2017, 37 (8): 2234-2239.)

[18] Jin W, Jung J H, Kang U, *et al.* Supervised and extended restart in random walks for ranking and link prediction in networks. [J]. PloS one, 2019, 14 (3): 1-23.

[19] Zhou Yinzuo, Wu Chencheng, Tan Lulu. Biased random walk with restart for link prediction with graph embedding method [J]. Physica A: Statistical Mechanics and its Applications, 2021 (6): 125783.

[20] Berahmand K, Nasiri E, Forouzandeh S, *et al.* A preference random walk algorithm for link prediction through mutual influence nodes in complex networks [J]. Journal of King Saud University-Computer and Information Sciences, 2021 (3) .

[21] Elahe N, Kamal B, Li Y F. A new link prediction in multiplex networks using topologically biased random walks [J]. Chaos, Solitons & Fractals, 2021 (151) .

[22] Vázquez A, Moreno Y. Resilience to damage of graphs with degree correlations [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2003, 67 (1): 015101.

[23] Wang Yan, Cao Xinxin, Weng Tongfeng, *et al.* Lowest-degree preference random walks on complex networks [J]. Physica A: Statistical Mechanics and its Applications, 2021, 577: 126075.

[24] Langville A N, Meyer C D. Google's pagerank and beyond: the science of search engine rankings [J]. The Mathematical Intelligencer, 2011, 30 (1): 68-69.

[25] 吕亚楠，韩华，贾承丰，等. 基于有偏向的重启随机游走链路预测算法 [J]. 复杂系统与复杂性科学, 2018, 15 (4): 17-24. (Lyu Yanan, Han Hua, Jia Chengfeng, *et al.* Link precdiction algorithm based on biased random walk with restart [J]. Complex Systems and Complexity Science, 2018, 15 (4): 17-24.)

[26] Fronczak A, Fronczak P. Biased random walks in complex networks: the role of local navigation rules [J]. Physical Review E, 2009, 80 (1): 016107.

[27] 徐全智. 随机过程及应用 [M]. 北京：高等教育出版社, 2013: 113-219. (Xu Quanzhi. Stochastic processes with its applications [M]. Beijing: Higher Education Press, 2013: 113-219.)

[28] Kim T H, Lee K M, Lee S. U. Generative image segmentation using random walks with estart [J]. Lecture Notes in Computer Science, 2008, 5304 (1): 264-275.

[29] 郑伟，王朝坤，刘璋，等. 一种基于随机游走模型的多标签分类算法 [J]. 计算机学报, 2010, 33 (8): 1418-1426. (Zheng Wei, Wang Chaokun, Liu Zhang, *et al.* Amulti-label classification algorithm based on random walk model [J]. Chinese Journal of Computers, 2010, 33 (8): 1418-1426.)

[30] Hanley J A, Mcneil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve [J]. Radiology, 1982, 143 (1): 29–36.

[31] Lawera M. Predictive Inference: an introduction [J]. Technometrics, 1995, 37 (1): 121–121.

[32] Kunejis J. Konect: the Koblenz network collection [C]// International conference on World Wide Web companion. Brazil: ACM Press, 2013: 1343-1350.